Methods for barcode analysis in optical DNA mapping

Dvirnas, Albertas

2021

Methods for barcode analysis in optical DNA mapping

ALBERTAS DVIRNAS

ASTRONOMY AND THEORETICAL PHYSICS | FACULTY OF SCIENCE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Science
Department of Astronomy and Theoretical Physics

ISBN 978-91-8039-125-2

9 789180 391252

Methods for barcode analysis in optical DNA mapping

Methods for barcode analysis in optical DNA mapping

by Albertas Dvirnas

LUND UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisor: Tobias Ambjörnsson
Faculty opponent: Jonas Nyvold Pedersen

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in the Lundmarkssalen at the Department of Astronomy and Theoretical Physics on Tuesday, the 25th of January 2022 at 10:00.

Organization
LUND UNIVERSITY

Department of Astronomy and Theoretical Physics
Sölvegatan 14A,
SE-223 62 Lund
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2022-01-25

Sponsoring organization

Author(s)
Albertas Dvirnas

Title and subtitle
Methods for barcode analysis in optical DNA mapping

Abstract

This thesis is composed of six papers, which all concern different methods and tools used for the analysis of barcodes in nanochannel-based Optical DNA Mapping (ODM). The first four papers consider densely-labeled barcodes while the last two consider sparsely-labeled barcodes.

Paper I presents a combinatorial auction algorithm for contig assembly ussing ODM barcodes as scaffolds.

Paper II concerns mapping of ODM barcodes on the human genome.

Paper III deals with bacterial typing.

Paper IV solves structural variation detection problem for competitive binding barcodes using Hidden Markov Models.

Paper V proposes the use of Sliding Frank-Wolfe methods for sparse-labeled single-frame ODM.

Paper VI extends the Sliding Frank-Wolfe methods from the analysis of single frame barcodes to multi-frame setting, where barcodes over multiple time-frames are averaged to improve the resolution.

Key words
DNA Barcoding, Optical Mapping, Computational Biology

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-8039-125-2 (print)
978-91-8039-126-9 (pdf)

Recipient's notes

Number of pages
219

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature ______________________ Date 2021-12-09

DOKUMENTDATABLAD enl SIS 61 41 21

Methods for barcode analysis in optical DNA mapping

by Albertas Dvirnas

LUND UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Funding information: This thesis was funded by the following grants:

- ”Non-equilibrium biophysics in nano-confined systems”, Grant number 2014-4305, The Swedish Research Council TA (Principal investigator, PI)
- The Erling Persson Family Foundation, ”A novel ultrafast method for detecting multiresistant bacteria in intensive care units”, 2017-2019 Fredrik Westerlund (PI), TA (co-PI)
- EU, ”BeyondSeq – Genomic diagnostics beyond the sequence”, 2015-2019, Yuval Ebenstein (PI)

Faculty of Science, Department of Astronomy and Theoretical Physics

ISBN: 978-91-8039-125-2 (print)
ISBN: 978-91-8039-126-9 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

To whom it may concern
Don't let the world bend
The formulas I write
They guide you far and wide

Contents

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Facilitated sequence assembly using densely labeled optical DNA barcodes: A combinatorial auction approach**

Albertas Dvirnas, Christoffer Pichler, Callum L Stewart, Saair Quaderi, Lena K Nyberg, Vilhelm Müller, Santosh Kumar Bikkarolla, Erik Kristiansson, Linus Sandegren, Fredrik Westerlund, Tobias Ambjörnsson
PloS one. 2018 Mar 9;13(3):e0193900

II **Enzyme-free optical DNA mapping of the human genome using competitive binding**

Vilhelm Müller, **Albertas Dvirnas**, John Andersson, Vandana Singh, Sriram Kk, Pegah Johansson, Yuval Ebenstein, Tobias Ambjörnsson, Fredrik Westerlund
Nucleic acids research. 2019 Sep 5;47(15):e89-

III **Cultivation-Free Typing of Bacteria Using Optical DNA Mapping**

Vilhelm Müller, My Nyblom, Anna Johnning, Marie Wrande, **Albertas Dvirnas**, Sriram Kk, Christian G Giske, Tobias Ambjörnsson, Linus Sandegren, Erik Kristiansson, Fredrik Westerlund
ACS infectious diseases. 2020 Apr 15;6(5):1076-84.

IV **Detection of structural variations in densely-labelled optical DNA barcodes: A hidden Markov model approach.**

Albertas Dvirnas, Callum Stewart, Vilhelm Müller, Santosh Kumar Bikkarolla, Karolin Frykholm, Linus Sandegren, Erik Kristiansson, Fredrik Westerlund, Tobias Ambjörnsson
Plos One, 16(11), e0259670. https://doi.org/10.1371/journal.pone.0259670

V **High-precision fluorophore localization in optical DNA mapping using the sliding Frank-Wolfe algorithm**

Albertas Dvirnas, Jonathan Jeffet, Yuval Ebenstein, Tobias Ambjörnsson
LU-TP 21-51

vi **Using time-averaging to increase precision in fluorophore localization in nanochannel-based optical DNA mapping**

Albertas Dvirnas, Jonathan Jeffet, Hemant Kumar, Henrik Nordanger, Fredrik Westerlund,Tobias Ambjörnsson
LU-TP 21-52

Publications not included in this thesis:

Optical DNA mapping combined with Cas9-targeted resistance gene identification for rapid tracking of resistance plasmids in a neonatal intensive care unit outbreak

Santosh K Bikkarolla, Viveka Nordberg, Fredrika Rajer, Vilhelm Müller, Muhammad Humaun Kabir, Sriram Kk, **Albertas Dvirnas**, Tobias Ambjörnsson, Christian G Giske, Lars Navér, Linus Sandegren, Fredrik Westerlund
MBio. 2019 Jul 9;10(4):e00347-19

Popular summary

Imagine Optical DNA Mapping as a factory printing stickers with barcodes that people have to scan when they are buying things in the supermarket. But the products that are being sold are not tofu and avocados, but DNA molecules. The customers are quite picky and want to find out what the product contains before buying it. The list of ingredients (the DNA sequence) is usually available to read, but the customer does not want to read the whole list. Their shopping list already is tedious enough. The customer then comes up with an idea: ”Maybe the barcodes that the factory produce should contain the information I need?”.

Obsessed by the idea, the customer contacts the printing factory. There the customer finds out that indeed, different products can be labeled by what they contain. There are mainly two different approaches: one can label some specific ingredient (sparse-labeling) or by amounts of that specific ingredient on different parts of the product (dense-labeling). But the amount of products in the super-market is huge, and the labeling methods are not without inaccuracies and errors, so the customer contacts a PhD student to help him out with the analysis of the barcodes.

Tools from applied mathematics, computational biology and biophysics are important in the barcode analysis of the Optical DNA mapping data. The factory becomes a (combinatorial) auction when the PhD student wants to align a fragmented list of ingredients to the barcode. Secret paths and states (in a hidden Markov model) are uncovered while looking for differences (structural variations) between two different barcodes. Statistical physics and transfer matrices are used when trying to predict theoretical barcodes before receiving the products. Pearson cross correlation is calculated to find specific category that products belong to (bacterial species). Sparse optimization is employed to improve the resolution of the barcodes.

The publications in this thesis contain my various approaches to developing the theoretical and computational tools for solving the problems of barcode analysis.

Introduction

Optical DNA mapping (ODM) provides us with a variety of methods for visualizing sequence-specific patterns along stretched, single DNA molecules [1]. A DNA barcode is a fluorescence profile obtained by sequence-specific labeling of DNA molecules with fluorescent dyes, then recording the fluorescence using a microscopy based setup. The stretching is done usually either on glass [2] or in nanochannels [1] [3]. Multiple labeling techniques for producing sequence-specific patterns have been developed, based either on enzymatic labeling [4], DNA melt mapping [5], or competitive binding (CB) [6, 7].

The experimental scheme of ODM field was first pioneered by [8], who utilized restriction enzymes to cut the DNA molecules at specific sites and visualized the length of the resulting fragments in an agarose gel. More recently, the ODM has advanced significantly, and companies such as Opgen and Bionano Genomics commercially released different platforms for optical DNA mapping. Bionano Genomics has commercially released two platforms Irys and Saphyr. These platforms are using labels produced by nicking enzymes and DNA stretched and visualized in nano-channels. The optical mapping is being increasingly used in various different genomic applications due to its improved accuracy, decreasing cost, reads of up to several megabasepairs and high throughput.

The experiments that use CB-based labeling have been shown to be of use for various different applications. It has been shown how it can be used for fast identification of bacteria [9] and bacterial resistance plasmids [10]. When combined with CRISPR/Cas9, it can be used for identification of antibiotic resistance genes on single plasmid molecules [11]. CB-based barcodes can also be used for detecting highly repetitive structures [12]. DNA melt mapping of single cell DNA genome combined with sequencing of short sequences was proposed as a complementary tools for the analysis of tumors [13]. Enzymatic labeling was used for identification of specific molecules of interest in genomes containing gigabasepairs [14]. Possibility of improving the resolution of the labels and reaching super-resolution was considered [15]. Enzymatic labeling approach is also used for producing optical maps, i.e. the locations of enzymes along the DNA as opposed to a sequence-specific pattern, and is

[1] All papers in this thesis consider DNA stretched in nanochannels.

used for structural variation analysis and sequence assembly [16, 17].

There are a number of mathematical challenges in the barcode analysis that have been addressed. For dealing with in-silico predictions, a competitive binding based approach was developed in [9]. Over multiple time-frames, a fast and scalable kymograph alignment method was developed in [18]. Hierarchical clustering was used to create a consensus of multiple barcodes of the same plasmid [10]. Attempts to reduce noise in the kymographs were considered in [19]. Re-sampling statistics was used to identify bacterial species [20]. Polymer physics of DNA stretched in nanochannels was extensively considered [21, 22], and relaxation of internal segments of DNA were considered using a simple Rouse-like model [23].

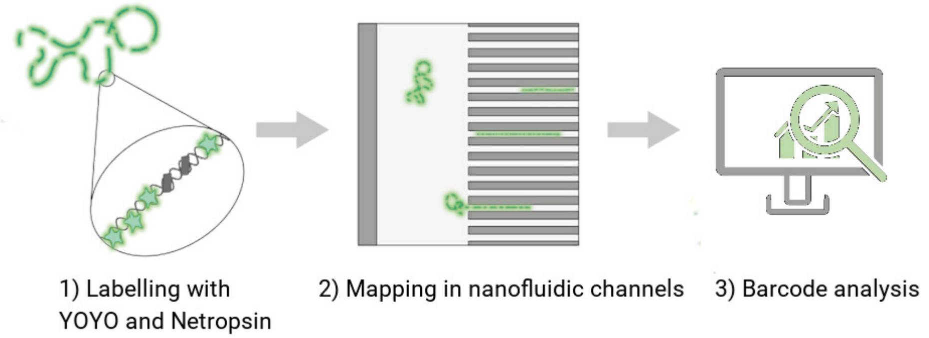

Figure 1: **Schematic overview of an experiment that produces a barcode as an output.** 1) DNA is fluorescently labeled, 2) DNA is analyzed in nanofluidic channels using fluorescence microscopy, 3) barcode analysis methods are developed to for instance assign the barcodes to the correct bacteria or identify the presence of resistance genes. The figure is adapted from Paper II.

Overall the field of ODM has been rapidly growing through-out my studies [24, 25, 26]. The emergence of commercial and high-throughput devices [27, 28, 12] has created access to huge amounts of data related to ODM. This motivates the development of new methodologies and tools to analyze the outputs of the ODM.

I have contributed to the field with developing computational tools and methods for a data-driven approach on CB-based optical DNA mapping (Papers I-IV) and enzymatic-labeling based data (Papers V-VI). My contribution to the different papers is summarized in Sec. 11. My contribution to the computational tools is described in Sec. 8. Other projects I had some smaller part in are summarized in Sec. 10. The main projects I have worked on was to develop a computational auction based contig assembly approach (Paper I), develop efficient and scalable tools for barcode matching to human genome (Paper II) and bacterial chromosomes (Paper III), introduce a pipeline for structural variations detection (Paper IV), a tool for high precision fluorophore localization (Paper V), and improving of localization accuracy using multi-timeframes (Paper VI).

The rest of the thesis is divided into two parts. The first part is introductory and gives a theoretical and conceptual background of tools used in my research, as well as individual contributions to each project, and the main text consists of full texts and supplementaries of the four published papers and two unpublished manuscripts.

1 DNA sequence as a barcode

In this thesis, we consider DNA molecules that have been labeled [7], stretched inside nanochannels [5] and imaged with a high sensitivity camera on a fluorescence microscope. Depending on the labeling method, the sequence-specific fluorescence patterns we get as an output of the imaging can be ”sparse” (meaning that we can visually separate individual peaks corresponding to labels), ”dense” (when we can't separate individual labels), and, in reality, there is also a gray zone between the two types of labels. See Fig. 2 for examples of typical fluorescence patterns.

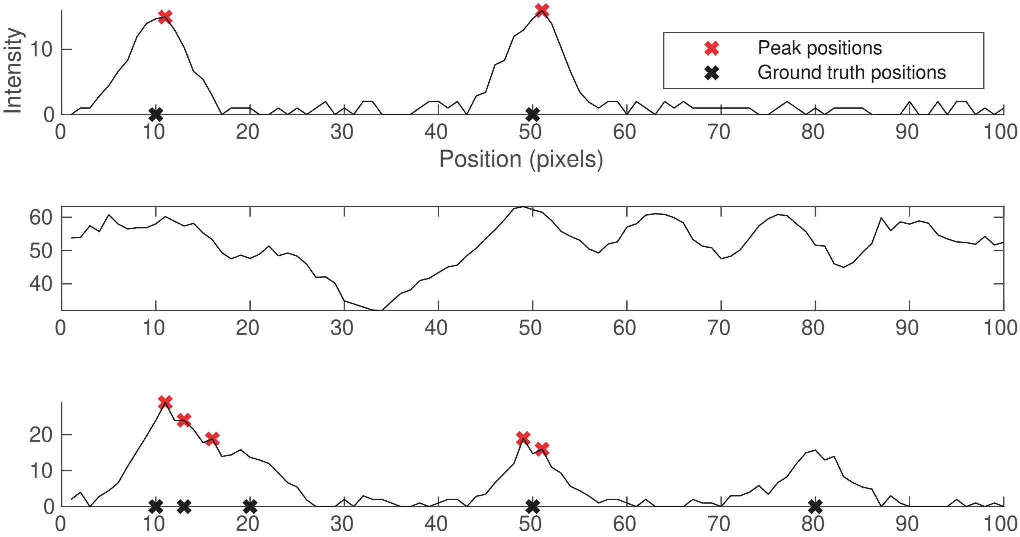

Figure 2: **Examples of typical barcodes representing different fluorescence patterns** (Top) Sparse labels, where peak positions closely match the ground truth positions (Middle) Dense labels, where the shape of the barcode is considered rather than individual label positions (Bottom) A hard example where some of the fluorophore locations are too close to each other to be resolved using simple peak finding methods.

Mathematically, we consider three alternative ways how to describe the different fluorescence patterns. The first one is to consider just the locations and their amplitudes and is used to represent optical maps, the second one combines locations (and possibly their amplitudes) into a measure, and the third introduces the definition of a barcode. In the first approach, one typically considers only the locations [29, 30], which leads to the definition of an *optical map*:

Definition 1 *Given detected locations of fluorophores along the DNA molecule $\theta = \theta_1, \ldots, \theta_n$ and their amplitudes $\mathbf{w} = \{w_1, \ldots, w_n\}$, the optical map is defined as the locations of fluorophores $\theta = \theta_1, \ldots, \theta_n$.*

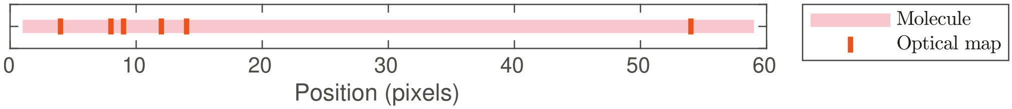

Figure 3: **Example of an optical map.** Only the locations of the six fluorophores $\theta = \theta_1, \ldots, \theta_6$ along the molecule are visualized with the optical map.

This representation is a typical result of the post-processing (peak detection) of the molecule snapshot on the output of "sparse" labeling fluorescence imaging experiments (see Fig. 2 (Top)), and is used in the optical map alignment. Typically, the peaks in a snapshot represent the positions of cuts or enzyme locations [31].

A second approach is to combine the locations θ and amplitudes $\mathbf{w}$ into a weighted Dirac mass sum of the form

$$m_{\mathbf{w},\theta} = \sum_{i=1}^{n} w_i \delta_{\theta_i} \tag{1}$$

This approach is used when defining a problem of recovery of spikes [32], and we adapt it to ODM in Papers V and VI.

The third representation is a standard approach for the densely-labeled experiments, and does not consider individual positions of the fluorophores [10, 33], but a sequence-specific finite array of numbers, called intensity profile or a barcode (a barcode is also analogous to a time series in time-series analysis), and defined as

Definition 2 *A barcode $\mathbf{B} \in \mathbb{R}^n$ is an array of real valued numbers $b_i \in \mathbb{R}^n$, i.e. $\mathbf{B} = [b_1, \ldots, b_n]$, where n is the length of $\mathbf{B}$.*

The barcode value b_i at position i represents photon count (including camera read out effects [34]) collected at a pixel i along the DNA molecule (i.e. positions in a nanochannel) in an optical DNA mapping (ODM) experiment. A few approaches also started using barcode approach for analysis of the optical maps [24, 35], i.e. also using the amplitudes $\mathbf{w}$ to improve the alignment against reference maps.

Finally, when the experiments are taken over a number of time-frames, then the barcodes from different time-frames are stacked together into a kymograph. See Fig. 4 for an example of a typical experimental kymograph.

Definition 3 *A kymograph $K \in \mathbb{R}^{n \times m}$ is a matrix where each row contains a barcode $\mathbf{B}_i$, $i = 1, \ldots m$.*

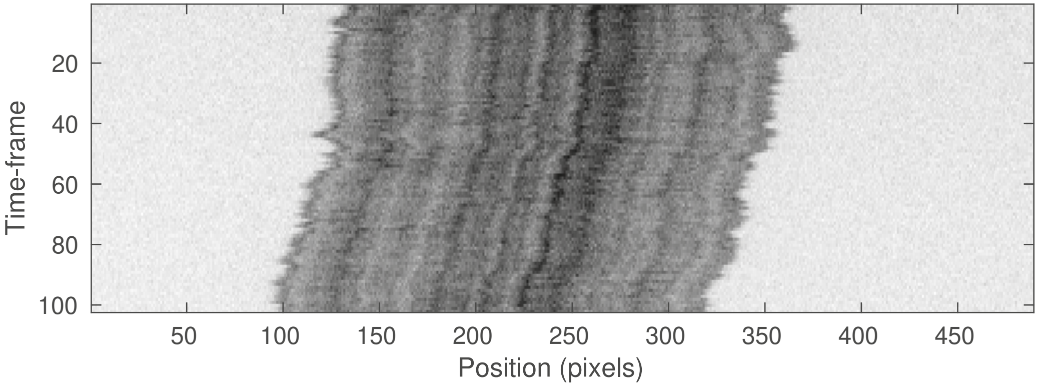

Figure 4: **Example of a typical experimental kymograph.** DNA molecule is labeled, inserted into a nanochannel for stretching, and then the barcodes (intensity profiles) are recovered at different times. A hundred barcodes from different time-frames are then stacked one on top of the other into a matrix. The kymograph is typically around 200-300 pixels long (this would correspond to 100-150 kilo base-pairs), and its center-of-mass is fluctuating over time. Intensity in the z-axis is inverted for clarity, so that the dark pixels represent high intensity.

Barcodes and kymographs are used in papers I-IV.

2 Binding probabilities of fluorescent dyes

For the densely-labeled ODM experiments, we use competitive binding with Netropsin and YOYO-1 to create a sequence-specific barcode [7]. Netropsin is a minor groove binder with an affinity to AT-rich regions. The binding of Netropsin blocks the binding of YOYO-1 (which is a non-selective, fluorescent intercalating dye).

To theoretically predict the CB-barcodes based on underlying DNA sequence, a few different biophysical formulations of 1-D lattice models could be used [36]. A standard approach is to use a transfer-matrix formulation. In this formulation the binding of YOYO-1 and Netropsin are described by a position dependent 8×8 transfer matrix $T(i)$. This transfer matrix includes as parameters the equilibrium binding constants for YOYO-1, and sequence-specific binding constants for Netropsin, and the free concentrations of YOYO-1 and netropsin. One then writes the total partition function Z as

$$Z = \mathbf{v}(1)^T \cdot T(1) \cdots T(N) \cdot \mathbf{v}(N+1) \tag{2}$$

where vectors $\mathbf{v}(1)$ and $\mathbf{v}(N+1)$ describe the boundary states at the ends of the DNA sequence [9]. One then also calculates a sequence dependent number of allowed states for YOYO-1

$$Z_{YOYO}(i) = \mathbf{v}(1)^T \cdot T(1) \cdots T(i-1) \cdot O_{YOYO} \cdot T(i) \cdots T(N) \cdot \mathbf{v}(N+1) \tag{3}$$

where O_{YOYO} is the projection operator that projects only to the states associated to $YOYO-1$. We then write the probability of YOYO-1 binding to a base-pair i as

$$p_{YOYO}(i) = \frac{Z_{YOYO}(i)}{Z} \tag{4}$$

See Fig. 5 for an example of the binding probability for a particular sequence.

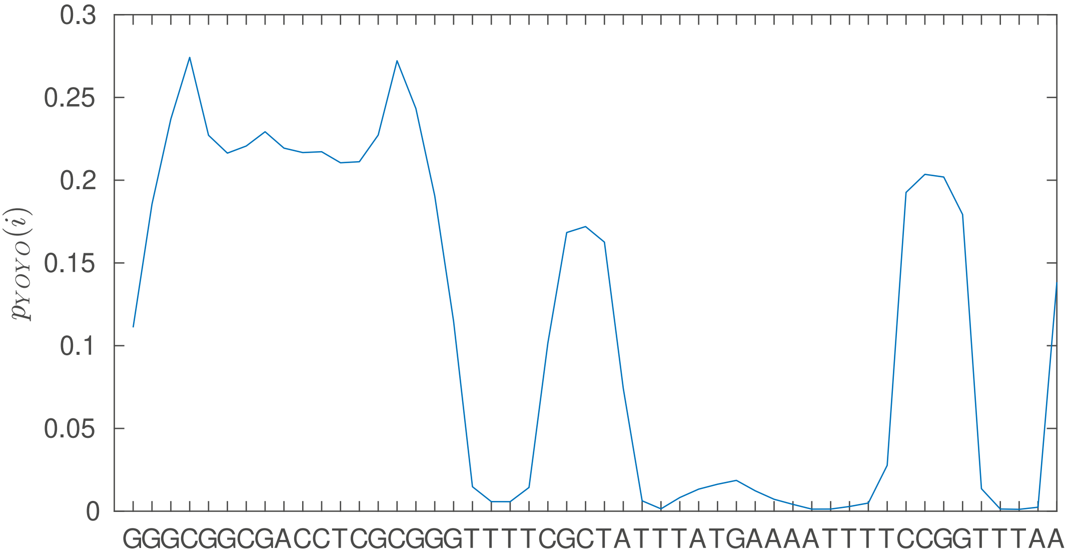

Figure 5: **Binding probabilities of YOYO-1 along a DNA sequence.** Zoomed-in version of the first 50 base-pairs binding probabilities.

In Paper I, we more accurately determined the YOYO-1 binding constant and the sequence-specific Netropsin binding parameters in the transfer matrix by training on an experimental barcode data-set of circular consensus barcodes from bacterial plasmids.

3 From binding probabilities to theoretical barcodes

The binding probability for a base-pair i describes how likely it is that a given fluorophore (YOYO-1 or any other fluorophore, depending on the experimental scheme) would be bound to that particular base-pair of the DNA. However, experiments have a point spread function (PSF) H that blurs the observations of fluorophores. Consider an array of M pixels, located at positions y_m (m denotes the index of the pixel, usually $y_1 = n_{bp}, y_2 = 2n_{bp}, \ldots, y_M = Mn_{bp}$, where n_{bp} is the number of base-pairs per pixel). Assuming that we are given binding probabilities $p(i)$ at base-pair positions $i = 1 \ldots N$, and a PSF

$$H(x) = e^{-x^2/(2\sigma^2)}, \tag{5}$$

then the theoretical barcode is given by

$$B(k) = \frac{1}{n_{bp}} \sum_{j=k-n_{bp}/2}^{k+n_{bp}/2} \left(\sum_{i=1}^{N} p(i) H(i-j) \right), \tag{6}$$

where $k = n_{bp}/2, 1 + n_{bp}/2, \ldots, K + n_{bp}/2$. See Fig.6 for an example of a theoretical barcode.

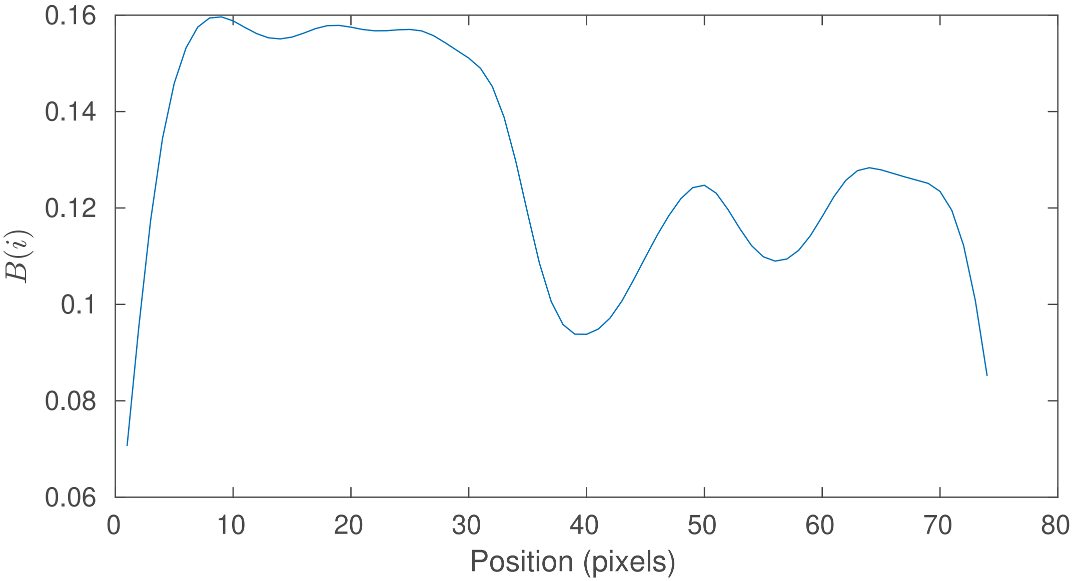

Figure 6: **An example of a theoretical barcode of *lambda phage*** . This barcode has a brighter and a darker region which is easily seen visibly. *lambda phage* is routinely used for determining the DNA extension in ODM experiments.

In paper II, we develop a scalable method to compute theoretical barcodes for large genomes. The computation can be made more efficient because PSF around each base-pair has only a local effect (a few kilo-basepairs). Using this information, the long genome can be divided into smaller overlapping parts and calculations done on each part independently.

4 Alignment scores

In many applications, such as creating consensus barcodes out of barcodes for individual molecules [10], there is a need to compare barcodes to each other. Also, there is a need of comparing the output of different processing methods. To that end, one needs to use alignment scores.

A lot of different alignment scores can be used in the analysis of optical maps [37] and DNA barcodes [13, 33]. The two standard approaches for DNA barcodes is to use a χ_i^2 function as in [13],

$$\chi_i^2 = \frac{1}{2N}\sum_{j=1}^{N}[\tilde{X}_j - \tilde{Y}_{i+j}]^2 \tag{7}$$

where $\tilde{X}$ is normalized intensities (subtracted by the mean and divided by standard deviation) for $j = 1 \dots N$ on the experimental barcode, and Y_{i+j} corresponding normalized theoretical barcode values. Slightly different figure-of-merit appears in previous works on densely-labeled ODM barcodes [10], where Pearson Cross Correlation (PCC) is used instead:

$$C_i = \frac{(X - \bar{X}) \cdot (Y^i - \bar{Y}^i)}{\sqrt{((X - \bar{X}) \cdot (X - \bar{X}))((Y^i - \bar{Y}^i) \cdot (Y^i - \bar{Y}^i))}} \tag{8}$$

where $Y^i = \{Y_i, Y_{i+1}, \dots, Y_{i+N-1}\}$ and $\bar{X}$ and $\bar{Y}^i$ denote the corresponding means of the barcodes, and the dots indicate scalar (dot) products. The version of PCC that we use in our papers is inspired by [38]. Both of these approaches relate to more standard method in time series analysis, the Euclidean Distance [39]

$$C_i = 1 - \frac{ED^2}{2N},\ \chi_i^2 = 1 - C_i \tag{9}$$

A few other methods that have not been used before in the analysis of barcodes but we have found to be of use are the Matrix Profile [40] and dynamical time warping (DTW) [41]. We introduced the use of matrix profile in Paper IV. This method compared sub-barcodes of one barcode to the sub-barcodes of another barcode, and is useful when for example part of the barcode is badly labeled. The DTW method elastically aligns two barcodes and can be used to improve alignment of barcodes with local fluctuations.

In case of optical maps, a standard approach is to use dynamical programming for aligning [31]. Alternatives include using a summed squared difference of the positions (that are linked by solving a linear assignment problem [42]), or PCC for the raw/recovered intensity profiles.

5 Evaluation metrics

When the ground truth is known, the accuracy of the alignment between two sets of points can be evaluated using Jaccard index (Jac), Recall (Rec), or Precision (Pre). Given some tolerance radius $r > 0$, the points along the estimated optical map are paired with the ground truth points if the distance is less than this radius. Paired fluorophores are called true positives (TP), un-paired estimated points are called false positives (FP), and un-paired ground truth points are called false negatives (FN). These quantities are then used to define the metrics

$$Jac = \frac{\#TP}{\#TP + \#FN + \#FP}, \; Rec = \frac{\#TP}{\#TP + \#FN}, \; Pre = \frac{\#TP}{\#TP + \#FP}. \tag{10}$$

Jaccard index measures the overall performance of detection by giving a measure of similarity between the two sets of fluorophores. The Recall and Precision metrics are used to measure the ability of the algorithm to minimize the FN and FP. When the points being compared are fluorophore positions, the fluorophores that are identified as TP are used to calculate the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_{TP}} (\theta_{i,TP} - \hat{\theta}_{i,TP})^2}{N_{TP}}} \tag{11}$$

where $\hat{\theta}_{i,TP}$ is the predicted location of the TP fluorophore i.

In paper III we use the evaluation metrics when we investigate whether CB-based ODM experiments on fragmented DNA molecules from patient samples can be used to decipher the type of bacteria present in the sample. For the samples with known bacteria species, we then identify as true positives (and call them discriminative) the DNA molecules that match to the correct bacterial species, and as false negatives the DNA molecules that match to an incorrect bacterial species. In Paper IV, we use the Rec metric when we estimate how many pixels of an ODM experiment the structural variation detection algorithm aligns correctly. In papers V and VI we used Jac index, Rec and RMSE when considering the accuracy of fluorophore detection.

6 Evaluating success of alignment

After calculating one of the alignment scores C_{align}, it is usually important also to evaluate the significance of the score as compared to a ”null model” (i.e. PDF distribution for maximum scores for non-matching barcodes). Such an approach is based on calculating p-values with respect to a null model, and requires no ground truth. If a null model probability density function (PDF) $f(C)$ is known for a score C, the cumulative distribution for the maximum $\hat{C}$ of uncorrelated scores $C_1, \ldots, C_\lambda$ is given as

$$\rho(\hat{C}) = \left(\int_{-\infty}^{\hat{C}} f(C) d(C) \right)^{\lambda} \tag{12}$$

The derivative of Eq. 12 with respect to $\hat{C}$ gives us a PDF for the maximum score. In practice the distribution depends on some fitting parameters which depend on the lengths of the two barcodes that are being aligned. In particular, λ is an effective number and has to be fitted. To determine the parameters, we generate 1000 random maximum scores C_i. The distribution of these scores is then fitted to the null model. The p-value is then estimated as $1 - CDF(C_{align})$, where CDF is the cumulative distribution of the null model with the fitted parameters. This gives us a statistics approach to estimating the significance of an alignment. in Paper I we have developed an approach based on this idea for Pearson Cross Correlation, since exact distribution is known in that case. We also use this approach in Papers II, and IV.

7 Polymer physics of DNA

The method of DNA stretching in all of the papers in this thesis is that of nanochannel confinement. Studying DNA at larger length scales, DNA can be modeled as a large polymer, built up by many mono-metric sub-units, parameterized by such parameters as size, shape, and molar mass. The extension of the DNA depends on the dimensions of the nanochannel, as well as the surrounding solvent, and the molar mass is determined by the number of monomers [43]. If the channels are smaller than the persistence length, the DNA molecule behaves according to the theory developed by Odijk [44]. The DNA cannot coil up and movement is restricted. For nanofluidic experiments, the channel dimensions are usually similar to the persistence length, resulting in transition regime in between the Odijk and de Gennes regime, what is commonly referred to as the extended de Genned regime. The degree of extensions is proportional to the contour length of the DNA, and therefore sequence specific information at some position along the sequence represents a similar position along the nanoconfined DNA molecule.

When in the nanochannel, the DNA is free to undergo center-of-mass diffusion as well as local conformation fluctuations. These fluctuations affect the barcode resolution. In order to understand and model these fluctuations several approaches have been made [45]. We have taken the approach of simulating Rouse-like chain of a number of beads connected by springs in paper VI, where we investigate how the effect of local fluctuations on the resolution in a sparsely-labeled ODM can be reduced by time-averaging.

8 Developing of computational tools for barcode analysis

Approximately half of the research time I spent through-out my PhD years was spent on developing computational tools and new features in the fields of image analysis, statistics, time series-type methods, comparison tools (combinatorial auction, hidden Markov model (HMM), pairwise dot matching, etc), super-resolution, optimization for my research group and for our collaborators. The necessity of having tools with graphical user interface arose once fast and scalable algorithms started being developed for nanochannel-based optical DNA mappings [18] [10]. The developed tools are routinely used by a group of our experimental collaborators in Fredrik Westerlund's lab at Chalmers University. The list of all the tools and their part in each of my research outcomes is summarized in Table 1.

Table 1: **List of computational tools developed and curated.**

Name of Tool	Which paper it is used in
Contig Assembly (CA)	Paper I
Human Chromosome Assembly (HCA)	Paper II and Paper III
HMMSV	Paper IV
HPFL-ODM	Paper V and Paper VI
lldev	Paper I, II, III, [46]

The contig assembly (CA) tool was used in the Paper I. This tool takes as an input a set of contig sequences and an experimental consensus barcode. The contig sequences are converted to in-silico theoretical barcodes, and the placement scores along the experimental consensus barcode are calculated. These scores are then converted to p-value based scores using a functional form of an extreme value distribution, and the contig barcodes that pass a p-value threshold are placed along the experimental consensus barcode using a computational auction algorithm. As the output the tool visualized the placement of the contig barcodes together with the placement positions. The theoretical challenge here was that, by the way of construction, the contigs should not overlap. This was why there was a need to introduce a combinatorial algorithm.

The HCA tool runs through the pipeline steps of comparing an experimental barcode against a set of in-silico theoretical barcodes obtained from reference DNA sequences. It combines in-silico theory prediction calculation, alignment of kymographs, calculation of alignment scores, and various output plots, and p-value calculation. This tool is made publicly available through the publication of this thesis [47].

The HMMSV tool is used for detecting structural variations between two barcodes (either experimental or in-silico). This tool runs a hidden markov model based alignment between the two barcodes. It also uses a Matrix Profile based p-value generation. This tool is made publically available together with Paper IV.

The HPFL-ODM tool detects and barcodes DNA molecules in pairs of fluorescence images, where one of the image-type represents a dense-labeling of the DNA backbone (for molecule detection) and the other image-type corresponds to sparsely-labeled barcode (for barcoding purposes). Here, the DNA molecules are labeled by two types of fluorescent dye molecules. The fluorescence of the two types of labels emit light at different wavelengths, so by filtering one can record both channels, the YOYO-1 channel for detecting the molecules and the other channel for sequence-specific barcode analysis. It then uses a Sliding-Frank-Wolfe-based approach to detect the positions of fluorophores. In case of multi-timeframe data, the fluorophore positions are tracked over the time-frames, and the linked trajectories are aligned and averaged to get a more accurate map.

The LLDEV tools combines many different tools for competitive-binding, melt-mapping, analysis of experimental barcodes. I have been curating this research tools package since the beginning of my PhD studies, and it is made publicly available, through the publication of this thesis, at [48].

9 Software tools from other projects

The number of different tools that are of interest to the barcode analysis of the Optical DNA mapping is growing rapidly, and here we briefly describe the potential use of some tools in the ODM analysis.

In Paper IV I introduced the use of methods from time-series analysis to barcode analysis. One particular method, VALMOD [49] is developed to extract variable length motifs in data series. This is relevant to the ODM as we could consider structural variations (i.e. sub-barcodes that represent insertions/translocations, inversions, etc.) as the data series. This is a possible future project related to the Paper IV.

The dynamical time warping method for time series sub-sequence search has an efficient implementation, Trillion [41], which I have adapted for the use in ODM, and it has potential use to increase the accuracy of the methods developed in Paper III and Paper IV.

Faster algorithms for calculating alignment scores can be developed and current methods optimized. Pearson Cross Correlation based scores are analyzed with an algorithm called SEC-C [50] and this could be of use in improving the speed of algorithms.

A promising open-source tool called OptiScan for analyzing sparse-labeled barcodes [35] could be used in combination or improved by adapting methods from Paper V and Paper VI. The results in Paper V were compared to those from OptiScan.

Ideas from topological data analysis could be applied for example for finding periodicity in the ODM data, as it is done for time-series data using the SW1PerS method [51].

10 Related Projects

There was a number of projects that I helped to develop in a bigger or a smaller way, but didn't directly participate in publications. These were topics of research of fellow students in my research group. They considered chromosome assembly using hierarchical clustering [52, 53], dual labels [54], analysis of cutting rate in DNA damage experiments [55].

A few other projects I have picked up from previous students, and continued developing them and eventually turned them into publications. These were the structural variation detection first considered in [56] (that became paper IV), gene-id [57] (the alignment algorithm is used in Papers II-IV), and contig assembly [58] (the first project I worked on, which became paper I).

11 Overview of publications

Here, I briefly summarize each paper in the thesis and specify my individual contributions.

Paper I: Facilitated sequence assembly using densely labeled optical DNA barcodes: A combinatorial auction approach

Albertas Dvirnas, Christoffer Pichler, Callum L Stewart, Saair Quaderi, Lena K Nyberg, Vilhelm Müller, Santosh Kumar Bikkarolla, Erik Kristiansson, Linus Sandegren, Fredrik Westerlund, Tobias Ambjörnsson
PloS one. 2018 Mar 9;13(3):e0193900

My roles in this paper: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Christoffer Pichler first began the project as his master thesis. I then proposed and implemented a combinatorial auction approach to solve the problem. The manuscript was written mostly by me and Tobias Ambjörnsson, I generated results and figures, wrote the methods section and the supplementary, and contributed to writing the results and discussion sections.

Paper II: Enzyme-free optical DNA mapping of the human genome using competitive binding

Vilhelm Müller, **Albertas Dvirnas**, John Andersson, Vandana Singh, Sriram Kk, Pegah Johansson, Yuval Ebenstein, Tobias Ambjörnsson, Fredrik Westerlund
Nucleic acids research. 2019 Sep 5;47(15):e89-

This article started as a summer project for Jon Andersson (from experimental group of Fredrik Westerlund at the Chalmers University of Technology) and I was developing data analysis methods and a software tool (called HCA) at the same time. The manuscript was mainly written by Vilhelm Müller (from experimental group of Fredrik Westerlund), while I contributed to the writing and editing and also wrote the supplementary.

Paper III: Cultivation-Free Typing of Bacteria Using Optical DNA Mapping

Vilhelm Müller, My Nyblom, Anna Johnning, Marie Wrande, **Albertas Dvirnas**, Sriram Kk, Christian G Giske, Tobias Ambjörnsson, Linus Sandegren, Erik Kristiansson, Fredrik Westerlund
ACS infectious diseases. 2020 Apr 15;6(5):1076-84.

This paper was continuation of the methods developed for Paper II. I contributed in developing the software for the matching against the theoretical intensity profiles and writing the data analysis section of the article.

Paper IV: Detection of structural variations in densely-labelled optical DNA barcodes: A hidden Markov model approach.

Albertas Dvirnas, Callum Stewart, Vilhelm Müller, Santosh Kumar Bikkarolla, Karolin Frykholm, Linus Sandegren, Erik Kristiansson, Fredrik Westerlund, Tobias Ambjörnsson
Plos One, 16(11), e0259670. https://doi.org/10.1371/journal.pone.0259670

My roles in the paper: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. I developed the methods pipeline, wrote the freely available software package HMMSV for this project, and generated all the final result figures and wrote the manuscript with the help of Tobias Ambjörnsson.

Paper V: High-precision fluorophore localization in optical DNA mapping using the sliding Frank-Wolfe algorithm

Albertas Dvirnas, Jonathan Jeffet, Yuval Ebenstein, Tobias Ambjörnsson
LU-TP 21-51

This paper appears as a manuscript in the thesis. I wrote the pipeline for extracting molecules from Optical DNA Mapping (where the experiments were run by Jonathan Jeffet from experimental group of Yuval Ebenstein at Tel Aviv University. I came up with the idea for the project and developed the software tool, HPFL-ODM. I implemented the Sliding-Frank Wolfe algorithm, generated the results, and wrote the manuscript.

Paper VI: Using time-averaging to increase precision in fluorophore localization in nanochannel-based optical DNA mapping

Albertas Dvirnas, Jonathan Jeffet, Hemant Kumar, Henrik Nordanger, Fredrik Westerlund,Tobias Ambjörnsson
LU-TP 21-52

This paper appears as a manuscript in the thesis. The project for detecting and tracking fluorescent molecules in ODM was started several years ago with experimental data from Jonathan Jeffet. Hemant Kumar and Henrik Nordanger worked on initial implementations of different types of analysis methods. I took over the project in 2018. I implemented the SFW method for multi-frame analysis and used the Rouse-like chain model to run the simulations, developed the pipeline together with Tobias Ambjörnsson, and generated all the results and wrote the manuscript.

12 Acknowledgements

I have had tremendous support from a lot of friends and colleagues through-out the four years of my PhD, that I almost never felt abandoned. Whether it was spontaneous lunch in the garden outside my window, or study company in one of many libraries of Lund, or being kidnapped for a driving lesson. So much that I often found myself wanting to lock my office door so I could concentrate on work, since the intimidation by many different fikas or never-ending conversations about geeky science with my supervisor seemed to actually never end. But the work was slowly progressing, the distractions of friends became less frequent, or rather quite different, as due to obvious reasons they became calls over Discord to play continuous rounds of a random game of Papayoo.

I want to thank my supervisor Tobias for never-ending support and motivation. To Fredrik for being my bridge to the lab (those who run the experiments) people. To Vilhelm for all the lists and e-mail conversations longer than this thesis. To all the lab people: Sriram, Rick, Santosh, Jonathan, My, Yuval, Gaurav, Alma. To student researchers and office friends in my own group, Saair, Henrik, Erik C, Erik T, Magnus, Jens, Hemant, Arthur, Lisa, Callum, Christopher. To the rest of publication collaborators: Anna, Erik K, Linus, Christian, Marie, Lena, Karolin, Viveka, Muhammad, Fredrika, Lars. To all the other colleagues at the department.

To my friends: Pijus, thanks to whom I ended up in Lund. Pranas and Simona, who are still very good friends many years after our studies together. Vytautas for many things, but not his temper. Mantas for devilstone. Mindaugas, who runs faster than me. Lukrecija, for being all I want in a friend. Marie for being a good friend. Femi for football. Jelte for apple-crumble. Rieke, for being a clown-fish. Femke for being nice. Amer for being even nicer. Nath for music. Emma for secrets. Sujeeth, for honesty. Irem for poesindia. Nadja, who showed me that a pineapple can be both incredible and inevitable.

To my teachers. Vita, for making me good at math. Elvyra, making me want to be better. Eligijus, for getting me through my undergraduate years. Gediminas Simonas, for ”kvantikos seminarai”.

Finally thanks to my family, Monika, Genadijus, Violeta, and the rest of our big family tree.

13 References

[1] Michal Levy-Sakin and Yuval Ebenstein. Beyond sequencing: optical mapping of dna in the age of nanotechnology and nanoscopy. *Current opinion in biotechnology*, 24(4):690–698, 2013.

[2] Robert K Neely, Jochem Deen, and Johan Hofkens. Optical mapping of dna: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311, 2011.

[3] Jonas O Tegenfeldt, Christelle Prinz, Han Cao, Steven Chou, Walter W Reisner, Robert Riehn, Yan Mei Wang, Edward C Cox, James C Sturm, Pascal Silberzan, et al. The dynamics of genomic-length dna molecules in 100-nm channels. *Proceedings of the National Academy of Sciences*, 101(30):10979–10983, 2004.

[4] Kyubong Jo, Dalia M Dhingra, Theo Odijk, Juan J de Pablo, Michael D Graham, Rod Runnheim, Dan Forrest, and David C Schwartz. A single-molecule barcoding system using nanoslits for dna analysis. *Proceedings of the National Academy of Sciences*, 104(8):2673–2678, 2007.

[5] Fredrik Persson and Jonas O Tegenfeldt. Dna in nanochannels—directly visualizing genomic information. *Chemical Society Reviews*, 39(3):985–999, 2010.

[6] Vilhelm Müller and Fredrik Westerlund. Optical dna mapping in nanofluidic devices: principles and applications. *Lab on a Chip*, 17(4):579–590, 2017.

[7] Lena K Nyberg, Fredrik Persson, Johan Berg, Johanna Bergström, Emelie Fransson, Linnea Olsson, Moa Persson, Antti Stålnacke, Jens Wigenius, Jonas O Tegenfeldt, et al. A single-step competitive binding assay for mapping of single dna molecules. *Biochemical and biophysical research communications*, 417(1):404–408, 2012.

[8] David C Schwartz, Xiaojun Li, Luis I Hernandez, Satyadarshan P Ramnarain, Edward J Huff, and Yu-Ker Wang. Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262(5130):110–114, 1993.

[9] Adam N Nilsson, Gustav Emilsson, Lena K Nyberg, Charleston Noble, Liselott Svensson Stadler, Joachim Fritzsche, Edward RB Moore, Jonas O Tegenfeldt, Tobias Ambjörnsson, and Fredrik Westerlund. Competitive binding-based optical dna mapping for fast identification of bacteria-multi-ligand transfer matrix theory and experimental applications on escherichia coli. *Nucleic acids research*, 42(15):e118–e118, 2014.

[10] Lena K Nyberg, Saair Quaderi, Gustav Emilsson, Nahid Karami, Erik Lagerstedt, Vilhelm Müller, Charleston Noble, Susanna Hammarberg, Adam N Nilsson, Fei Sjöberg, et al. Rapid identification of intact bacterial resistance plasmids via optical mapping of single dna molecules. *Scientific reports*, 6(1):1–10, 2016.

[11] Vilhelm Müller, Fredrika Rajer, Karolin Frykholm, Lena K Nyberg, Saair Quaderi, Joachim Fritzsche, Erik Kristiansson, Tobias Ambjörnsson, Linus Sandegren, and Fredrik Westerlund. Direct identification of antibiotic resistance genes on single plasmid molecules using crispr/cas9 in combination with optical dna mapping. *Scientific reports*, 6(1):1–11, 2016.

[12] Franziska M Esmek, Tim Erichlandwehr, Dennis HB Mors, Manja Czech-Sioli, Marlin Therre, Thomas Günther, Adam Grundhoff, Nicole Fischer, and Irene Fernandez-Cuesta. Real time, in-line optical mapping of single molecules of dna. *Biosensors and Bioelectronics: X*, page 100087, 2021.

[13] Rodolphe Marie, Jonas N Pedersen, Loic Bærlocher, Kamila Koprowska, Marie Pødenphant, Céline Sabatel, Maksim Zalkovskij, Andrej Mironov, Brian Bilenberg, Neil Ashley, et al. Single-molecule dna-mapping and whole-genome sequencing of individual cells. *Proceedings of the National Academy of Sciences*, 115(44):11192–11197, 2018.

[14] Nathaniel O Wand, Darren A Smith, Andrew A Wilkinson, Ashleigh E Rushton, Stephen J W Busby, Iain B Styles, and Robert K Neely. Dna barcodes for rapid, whole genome, single-molecule analyses. *Nucleic acids research*, 47(12):e68–e68, 2019.

[15] Jonathan Jeffet, Asaf Kobo, Tianxiang Su, Assaf Grunwald, Ori Green, Adam N Nilsson, Eli Eisenberg, Tobias Ambjörnsson, Fredrik Westerlund, Elmar Weinhold, et al. Super-resolution genome mapping in silicon nanochannels. *ACS nano*, 10(11):9823–9830, 2016.

[16] Ernest T Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K Das, Michael D Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology*, 30(8):771–776, 2012.

[17] Siavash Raeisi Dehkordi, Jens Luebeck, and Vineet Bafna. Fandom: Fast nested distance-based seeding of optical maps. *Available at SSRN 3774512*.

[18] Charleston Noble, Adam N Nilsson, Camilla Freitag, Jason P Beech, Jonas O Tegenfeldt, and Tobias Ambjörnsson. A fast and scalable kymograph alignment algorithm for nanochannel-based optical dna mappings. *PloS one*, 10(4):e0121905, 2015.

[19] Paola C Torche, Vilhelm Müller, Fredrik Westerlund, and Tobias Ambjörnsson. Noise reduction in single time frame optical dna maps. *PloS one*, 12(6):e0179041, 2017.

[20] Arno Bouwens, Jochem Deen, Raffaele Vitale, Laurens D'Huys, Vince Goyvaerts, Adrien Descloux, Doortje Borrenberghs, Kristin Grussmayer, Tomas Lukes, Rafael Camacho, et al. Identifying microbial species by single-molecule dna optical mapping and resampling statistics. *NAR genomics and bioinformatics*, 2(1):lqz007, 2020.

[21] Walter Reisner, Keith J Morton, Robert Riehn, Yan Mei Wang, Zhaoning Yu, Michael Rosen, James C Sturm, Stephen Y Chou, Erwin Frey, and Robert H Austin. Statics and dynamics of single dna molecules confined in nanochannels. *Physical Review Letters*, 94(19):196101, 2005.

[22] Liang Dai, C Benjamin Renner, and Patrick S Doyle. The polymer physics of single dna confined in nanochannels. *Advances in colloid and interface science*, 232:80–100, 2016.

[23] Aashish Jain, Julian Sheats, Jeffrey G Reifenberger, Han Cao, and Kevin D Dorfman. Modeling the relaxation of internal dna segments during genome mapping in nanochannels. *Biomicrofluidics*, 10(5):054117, 2016.

[24] Jonathan Jeffet, Sapir Margalit, Yael Michaeli, and Yuval Ebenstein. Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale. *Essays in Biochemistry*, 65(1):51–66, 2021.

[25] Wahab A Khan and Diana M Toledo. Applications of optical genome mapping in next-generation cytogenetics and genomics. *Advances in Molecular Pathology*, 4:27–36, 2021.

[26] Yuxuan Yuan, Claire Yik-Lok Chung, and Ting-Fung Chan. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal*, 2020.

[27] Sven Bocklandt, Alex Hastie, and Han Cao. Bionano genome mapping: high-throughput, ultra-long molecule genome analysis system for precision genome assembly and haploid-resolved structural variation discovery. *Single Molecule and Single Cell Sequencing*, pages 97–118, 2019.

[28] Sriram Kk, Yii-Lih Lin, Tsegaye Sewunet, Marie Wrande, Linus Sandegren, Christian G Giske, and Fredrik Westerlund. A parallelized nanofluidic device for high-throughput optical dna mapping of bacterial plasmids. *Micromachines*, 12(10):1234, 2021.

[29] Alden King-Yung Leung, Tsz-Piu Kwok, Raymond Wan, Ming Xiao, Pui-Yan Kwok, Kevin Y Yip, and Ting-Fung Chan. Omblast: alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics*, 33(3):311–319, 2017.

[30] Le Li, Alden King-Yung Leung, Tsz-Piu Kwok, Yvonne YY Lai, Iris K Pang, Grace Tin-Yun Chung, Angel CY Mak, Annie Poon, Catherine Chu, Menglu Li, et al. Omsv enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome biology*, 18(1):1–19, 2017.

[31] Anton Valouev, Lei Li, Yu-Chi Liu, David C Schwartz, Yi Yang, Yu Zhang, and Michael S Waterman. Alignment of optical maps. *Journal of Computational Biology*, 13(2):442–462, 2006.

[32] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank–wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.

[33] Vilhelm Müller, Nahid Karami, Lena K Nyberg, Christoffer Pichler, Paola C Torche Pedreschi, Saair Quaderi, Joachim Fritzsche, Tobias Ambjörnsson, Christina Åhrén, and Fredrik Westerlund. Rapid tracing of resistance plasmids in a nosocomial outbreak using optical dna mapping. *ACS infectious diseases*, 2(5):322–328, 2016.

[34] Michael Hirsch, Richard J Wareham, Marisa L Martin-Fernandez, Michael P Hobson, and Daniel J Rolfe. A stochastic model for electron multiplication charge-coupled devices–from theory to practice. *PloS one*, 8(1):e53671, 2013.

[35] Mehmet Akdel, Henri Van De Geest, Elio Schijlen, Irma MH Van Rijswijck, Eddy J Smid, Gabino Sanchez-Perez, and Dick De Ridder. Signal-based optical map alignment. *PloS one*, 16(9):e0253102, 2021.

[36] Vladimir B Teif and Karsten Rippe. Calculating transcription factor binding maps for chromatin. *Briefings in bioinformatics*, 13(2):187–201, 2012.

[37] Daniel Sage, Thanh-An Pham, Hazen Babcock, Tomas Lukes, Thomas Pengo, Jerry Chao, Ramraj Velmurugan, Alex Herbert, Anurag Agrawal, Silvia Colabrese, et al. Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. *Nature methods*, 16(5):387–395, 2019.

[38] A Mueen, K Viswanathan, CK Gupta, and E Keogh. The fastest similarity search algorithm for time series subsequences under euclidean distance. url: www cs unm edu/ mueen. *FastestSimilaritySearch html (Accessed 24 May 2016)*, 2015.

[39] Abdullah Mueen, Hossein Hamooni, and Trilce Estrada. Time series join on subsequence correlation. In *2014 IEEE International Conference on Data Mining*, pages 450–459. IEEE, 2014.

[40] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1317–1322. Ieee, 2016.

[41] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and

mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270, 2012.

[42] Iain S Duff and Jacko Koster. On algorithms for permuting large entries to the diagonal of a sparse matrix. *SIAM Journal on Matrix Analysis and Applications*, 22(4):973–996, 2001.

[43] Walter Reisner, Jonas N Pedersen, and Robert H Austin. Dna confinement in nanochannels: physics and biological applications. *Reports on Progress in Physics*, 75(10):106601, 2012.

[44] Theo Odijk. Scaling theory of dna confined in nanochannels and nanoslits. *Physical Review E*, 77(6):060901, 2008.

[45] Douglas R Tree, Yanwei Wang, and Kevin D Dorfman. Modeling the relaxation time of dna confined in a nanochannel. *Biomicrofluidics*, 7(5):054118, 2013.

[46] Santosh K Bikkarolla, Viveka Nordberg, Fredrika Rajer, Vilhelm Müller, Muhammad Humaun Kabir, Sriram Kk, Albertas Dvirnas, Tobias Ambjörnsson, Christian G Giske, Lars Navér, et al. Optical dna mapping combined with cas9-targeted resistance gene identification for rapid tracking of resistance plasmids in a neonatal intensive care unit outbreak. *MBio*, 10(4):e00347–19, 2019.

[47] Hca 4.4, [10.5281/zenodo.5718183].

[48] lldev.v.0.5.3 [10.5281/zenodo.5718208].

[49] Michele Linardi, Yan Zhu, Themis Palpanas, and Eamonn Keogh. Matrix profile x: Valmod-scalable discovery of variable-length motifs in data series. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1053–1066, 2018.

[50] Nader Shakibay Senobari, Gareth J Funning, Eamonn Keogh, Yan Zhu, Chin-Chia Michael Yeh, Zachary Zimmerman, and Abdullah Mueen. Super-efficient cross-correlation (sec-c): A fast matched filtering code suitable for desktop computers. *Seismological Research Letters*, 90(1):322–334, 2019.

[51] Jose A Perea, Anastasia Deckard, Steve B Haase, and John Harer. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC bioinformatics*, 16(1):1–12, 2015.

[52] Erik Clarkson. Chromosomal dna barcode assembly using hierarchical clustering matrix method: Including elastic matching. *[Bachelor thesis]*, 2020.

[53] Wensi Zhu. Hierarchical clustering matrix (hcm) method applied to dna barcode assembly for bacterial chromosomes. *[Master thesis]*, 2018.

[54] Erik Torstensson, Gaurav Goyal, Anna Johnning, Fredrik Westerlund, and Tobias Ambjörnsson. Combining dense and sparse labeling in optical DNA mapping. *PLOS ONE*, 16(11):1–20, 2021.

[55] Magnus Brander. Dna damage–a novel method to measure the rate of single-stranded breaks from fragmenting dsdna in nanochannels–theory and modeling. *[Master thesis]*, 2020.

[56] Callum Stewart. Methods for structural variation detection and improved theory prediction for densely labeled dna barcodes. *[Master thesis]*, 2017.

[57] Henrik Nordanger. Gene-id using simultaneous dna barcoding and enzymatic labeling. *[Master thesis]*, 2017.

[58] Christoffer Pichler. Contig assembly and plasmid analysis using dna barcodes. *[Master thesis]*, 2016.